https://doi.org/10.5194/amt-2019-200
Preprint. Discussion started: 11 September 2019

Atmospheric Measurement Techniques Discussions

Total variation of atmospheric data: covariance minimization about objective functions to detect conditions of interest

Nicholas Hamilton
National Renewable Energy Laboratory, Golden, Colorado, USA
Correspondence: Nicholas Hamilton (nicholas.hamilton@nrel.gov)

Abstract. Identification of atmospheric conditions within a multivariable atmospheric data set is a necessary step in the validation of emerging and existing high-fidelity models used to simulate wind plant flows and operation. Most often, conditions of interest are determined as those that occur most frequently, given the need for well-converged statistics from observations against which model results are compared. Aggregation of observations without regard to covariance between time series discounts the dynamical nature of the atmosphere and is not sufficiently representative of wind plant operating conditions. Identification and characterization of continuous time periods with atmospheric conditions that have a high value for analysis or simulation sets the stage for more advanced model validation and the development of real-time control and operational strategies. The current work explores a single metric for variation of a multivariate data sample that quantifies variability within each channel as well as covariance between channels. The *total variation* is used to identify periods of interest that conform to desired objective functions, such as quiescent conditions, ramps or waves of wind speed, and changes in wind direction. The direct detection and classification of events or periods of interest within atmospheric data sets is vital to developing our understanding of wind plant response and to the formulation of forecasting and control models.

1 Introduction

Parsing multivariate data sets that are ever growing in size and complexity can be a daunting task for researchers seeking to identify periods or events of interest in time series data (Preston et al., 2009; Shahabi and Yan, 2003). This is especially true for wind energy research seeking to validate high-fidelity numerical models against field observations (Barthelmie et al., 2015; Larsen et al., 2013; Sørensen and Shen, 2002). Wind plants operate continuously over time periods spanning years and across a broad range of atmospheric conditions, each of which implicitly impact the operation of the wind plant, either in terms of power production, operations and maintenance costs, or energy forecasting for grid integration.

Field observations of wind plants are typically collected by instrumentation mounted to wind turbines or meteorological towers (met masts) and by supervisory control and data acquisition (SCADA) systems. Wind plant data sets typically include measurements of wind speed and direction, local temperature and pressure, and wind turbine operational data, such as operational status, power production, and nacelle position. Each of the atmospheric quantities of interest may be classified as nonhomogenous stochastic variables that are fundamentally connected (i.e. strongly interdependent).

https://doi.org/10.5194/amt-2019-200
Preprint. Discussion started: 11 September 2019

Atmospheric Measurement Techniques Discussions

Wind speed ramps are of particular interest in wind plant power forecasting due to the need to balance energy production against demand curves and in the planning of required reserves and base loads (Sevlian and Rajagopal, 2012; Zhang et al., 2014). Previous work has focused on forecasting of mesoscale wind speed acceleration (Bossavy et al., 2013; Ferreira et al., 2011), generally concentrating on risk and reliability issues for wind turbines. Ramp event detection has been a research focus for more than a decade, (Cutler et al., 2007; Ferreira et al., 2013; Hannesdóttir and Kelly, 2019), and has produced some specific recommendations for individual turbine controls and the influence on operations and maintenance costs or activities. Previous research in wind speed ramps is not easily generalized to the identification and characterization of other dynamical events of interest, despite parallels in the detection process and considerations for wind turbine or plant operations and controls.

Detection of events in noisy data is of particular interest in the case of turbulent atmospheric data sets, especially given the need for more sophisticated forecasting systems (Belušić and Mahrt, 2012; Fulcher, 2018; Gamage and Hagelberg, 1993; Kang et al., 2014, 2017; Sun et al., 2015). One of the more common event detection methods leverages the continuous or discrete wavelet transform (Gamage and Hagelberg, 1993; Kumar and Foufoula-Georgiou, 1997; Lilly, 2017). Wavelet transforms leverage time-frequency signals designed to have specific properties that make them easy to use in signal processing applications. However, wavelet transformation remains computationally intensive and requires a fair amount of expertise to implement effectively and avoid the common pitfalls of signal shift sensitivity and the poor representation of phase and directionality (Taswell, 2001). A more direct method simply considers the covariance matrix of the input data, which represents the statistical spread of each data channel as well as cross-correlated variability (Eaton, 1983; Wasserman, 2013). Reducing the variability of a sample of multi-dimensional observations to a single metric is a necessary step to using numerical methods such as least-squares minimization for event detection and classification.

Simultaneous observation of multiple thermodynamic and kinematic quantities reported by met masts are necessary to characterize the dynamical state of the atmosphere (Barthelmie et al., 2014; Hansen et al., 2012). Directly considering multiple disparate data channels simultaneously represents a challenge in that each quantity has different engineering units and that variation within each channel may occur over a distinct scale. Atmospheric conditions are frequently characterized by considering wind speed, wind direction, and turbulence intensity or thermal stability, each of which have different units, ranges, and statistical properties. Consideration of these data independently is likely to provide skewed or biased reports of variability and can offer only a limited range of insights as to the state of the atmosphere or dynamical events relevant to the operation of wind energy assets. Further, and perhaps most importantly, direct comparison of statistical quantities (measures of central tendency, variability, or higher statistical moments) discount the inherent coupling between quantities of interest that underpin atmospheric physics (Hannesdóttir and Kelly, 2019; Preston et al., 2009; Shahabi and Yan, 2003).

The following work explores an application of numerical analysis methods to atmospheric data to identify continuous periods of interest within met mast time series data. The source of the data and their treatment are discussed briefly, although the wind plant and met mast are not in themselves imperative to the demonstration of the method or its utility. A discussion of aggregate statistical measures of the data is followed by a formal definition of the total variability of a block of time series data, and applications using the total variation as a metric to identify specific dynamical events of interest. Finally, the method sensitivity to outliers is analyzed, ending with a discussion of broader applications and extensions to the method.

https://doi.org/10.5194/amt-2019-200
Preprint. Discussion started: 11 September 2019

Atmospheric Measurement Techniques Discussions

2 Data and quality control

Data used to demonstrate the current method for detecting conditions of interest issue from met mast signals at the Lillgrund Wind Farm, located 10 km off the coast of southern Sweden in the Kattegat Strait. Lillgrund is comprised of 48 Siemens SWT-2.3-93 wind turbines and has a rated nameplate capacity of 110 MW. The layout of the Lillgrund wind plant is shown in Fig. 1(a), where each turbine location is denoted with a marker whose color is representative of the average power produced over the time period analyzed below. Production data have been normalized to an interval of [0, 1], representing the wind turbines producing the least and greatest power, respectively. Operational data (SCADA, power production, turbine availability) from the wind farm are not discussed further in the following analysis, although a brief summary of future applications of the method is provided in the conclusions section, including thoughts on wind plant performance and SCADA data. Data used to demonstrate the calculation of total variation and identify periods of interest come from the met mast, located at the southwest corner of the wind plant, indicated in Fig. 1(a) with an open marker.

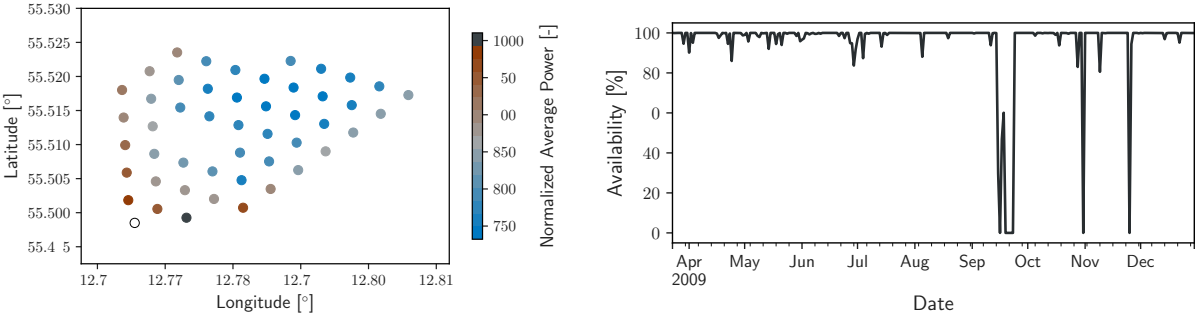
Within any wind plant data, particular conditions of interest are typically identified either by way of aggregate statistical metrics or by identifying "well-behaved" time periods exhibiting a dynamical event or atmospheric condition of interest. Kinematic and thermodynamic atmospheric quantities that are expected to have the greatest impact on the performance of a wind plant are the wind speed u, wind direction θ, and the atmospheric stability, considered either in an instantaneous or time-averaged sense. The stability of the atmosphere (typically quantified by the Monin–Obukhov stability parameter or the Richardson number) indicates the degree to which the energy equation is coupled to the Navier–Stokes equations and whether it represents either a source or sink of momentum (Kumar et al., 2006). Forcing in the momentum equations as indicated by the presence and sign of a buoyancy term is manifested in atmospheric flow as vertical turbulent mixing, and is an important overall factor in the energy balance relevant to wind plant operation. Thermal stability has a significant effect on atmospheric turbulence and the structure of wind turbine wakes, wake interaction, and thus the overall energy balance within the wind plant (Ali et al., 2019). In lieu of a time series of Richardson number or the Monin–Obukhov stability parameter, turbulence intensity (TI) is used in the current demonstration as a proxy for stability. While TI is an imperfect estimate of atmospheric stability from a fluid mechanical or atmospheric perspective, it does provide some sense of the energy contained in the fluctuating flow field, and is well-suited for presenting the utility of the total variation method below. Additionally, TI is a quantity frequently used in the wind energy community to characterize wind plant operating conditions and is often accessible through instrumentation on met masts or wind turbine nacelles making it an appropriate choice for the current demonstration.

Raw data used to demonstrate the current methods include high-frequency (20 Hz) observations of u and θ reported by the met mast between March and December 2009. Wind speed and direction data were binned to a temporal resolution of 1 min, from which mean and standard deviations were calculated. Turbulence intensity in each bin is estimated as the ratio of the retained 1-min statistics for wind speed as $TI = \sigma_u/u$. As with most field observations, data availability from each channel is less than 100%, as instruments require maintenance, loose connectivity to data acquisition systems, or shut down to prevent damage under certain conditions. Binning the data into 1-min periods smooths the observed time series of wind speed and direction, and reduces the noise reported by the cup anemometer and wind vane. Additional quality-control steps

https://doi.org/10.5194/amt-2019-200
Preprint. Discussion started: 11 September 2019

Atmospheric Measurement Techniques Discussions

for the data include omitting any 1-min period in which not all data channels are correctly reported (e.g. data are missing or report a single, fixed value) from further consideration. Any time stamp associated with wind speeds less than 1 m/s, when wind speed observations reported by cup anemometers and wind vanes are not considered to be reliable (IEC, 2005), are also removed from the data set. Fig. 1(b) shows data availability of the record as a percent of the total number of data possible per day. The final quality-control step implemented for the current study is to exclude data that are not part of any continuous set of observations of at least 60 min. The current method searches continuous data samples to identify atmospheric conditions and events of interest. Rather than infill or interpolate data, periods with missing values are simply excluded from consideration.



(a) Lillgrund layout, wind turbines are filled markers, met mast is open marker

(b) Daily data availability from met mast

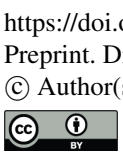
Figure 1. Wind turbines, met mast, and data availability from Lillgrund wind plant

3 Statistical view of atmospheric conditions

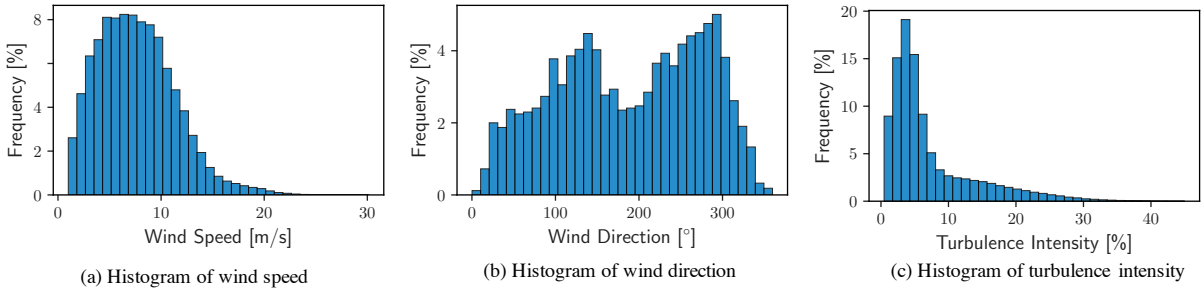
Characterization of the atmospheric conditions is most often pursued through aggregate statistics, that is without explicitly considering the evolution of atmospheric variables. Statistical moments (arithmetic mean values, variances, and higher-order statistical moments) may reflect the occurrence of infrequent events, but do not convey dynamical evolution of variables or their correlation in time. Histograms of each of the data channels are provided in Fig. 2, showing characteristic behavior for the wind speed and turbulence intensity distributions.

The wind direction (Fig. 2(b)) shows several key features typical of atmospheric records; first, it identifies the prevailing wind directions as per the number of observations within each direction sector (10°) and, second, it shows that virtually no observations correspond with wind directly out of the north. According to the International Electrotechnical Commission standard governing the placement and reliability of instrumentation (IEC, 2005), met masts should be placed sufficiently far from the nearest upstream obstacle, or risk introducing bias and increased uncertainty into the record. This limitation can be difficult or prohibitively expensive to accommodate due to logistical constraints, especially in offshore settings where placement is often strictly limited.

https://doi.org/10.5194/amt-2019-200
Preprint. Discussion started: 11 September 2019

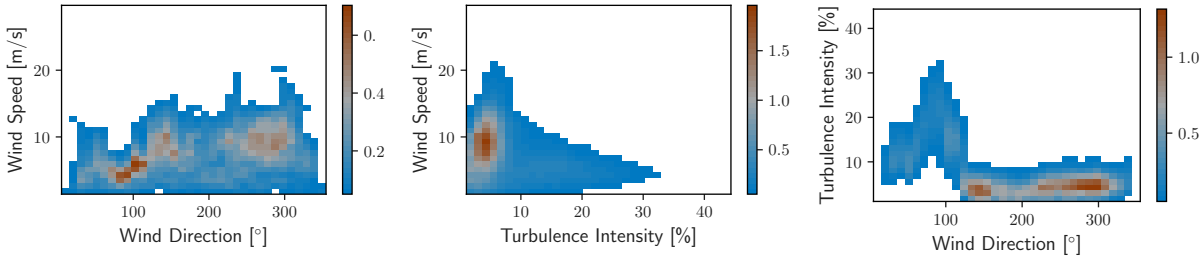


Atmospheric Measurement Techniques Discussions



(a) Histogram of wind speed (b) Histogram of wind direction (c) Histogram of turbulence intensity

Figure 2. Histograms of quality-controlled met mast data



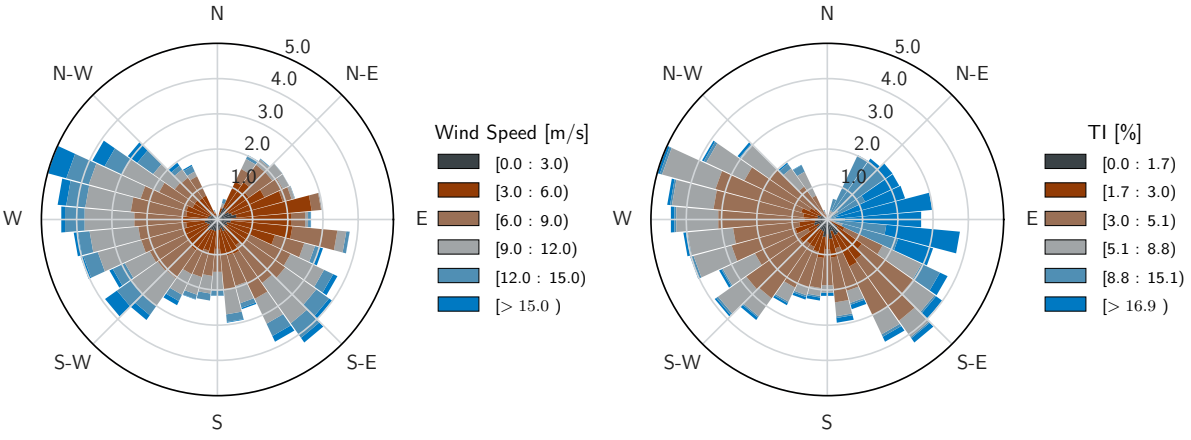
(a) Distribution of u and $\boldsymbol{\theta}$ (b) Distribution of u and TI (c) Distribution of TI and $\boldsymbol{\theta}$

Figure 3. Two-dimensional histograms of met mast data. Color information conveys percent of total observations for each pair of variable values.

Each of the histograms in Fig. 2 categorizes a single quantity without regard to the variation of the others; each single-variable histogram effectively integrates the available observations over the ranges of the other two variables. More complex treatment of the data is required to take into account the simultaneous variability of more than one channel. Fig. 3 shows two-dimensional histograms with two-way permutations of the data channels. The colorbar associated with each subfigure describes the frequency of observing atmospheric conditions within a given bin described by the respective variables. In each of the histograms, a threshold has been applied to the frequency of observations. Any bin representing less than 0.5% of the total observations has been filtered out to highlight more common conditions. Two-dimensional histograms demonstrate that the atmospheric conditions are far more complex than is possible to estimate from pairwise consideration of any two of the one-dimensional histograms in Fig. 2. This increase in complexity arises from the interdependence of each of the variables retained for analysis. An observation from the two-dimensional histograms that is not immediately evident in one-dimensional histograms is that the greatest turbulence intensity comes from a single, distinct sector of wind directions. Placement of the met mast with respect to the wind turbines contributes to a sharp increase of TI in the range of 15–45% and is not typical of unobstructed measurements. Reports of high TI likely result from the introduction of turbulence to the flow by the wind turbines or wind plant from directions between 70°–110°.

https://doi.org/10.5194/amt-2019-200
Preprint. Discussion started: 11 September 2019

Atmospheric Measurement Techniques Discussions



(a) Wind rose (b) *TI* rose

Figure 4. Wind (a) and TI (b) roses from met mast data

In wind energy research, the coupling of wind direction with either wind speed or turbulence intensity is often visualized by a rose diagram. Wind roses (and TI roses) contain the same information as the two-dimensional histograms from Fig. 4, but convey it on a polar projection representative of the compass, thus making them more intuitive to read for many users. Fig. 4 shows wind and TI roses for the considered data. The rose diagrams highlight directional dependence of the mapped variable. For example, Fig. 4(b) demonstrates that the greatest turbulence intensity is highly correlated with winds from the sector of 70°–110°. This is the range of directions in which the met mast is waked by the wind turbine located to the west.

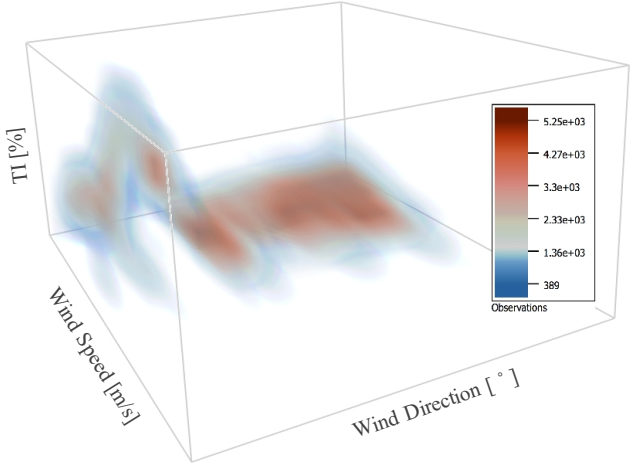


Figure 5. Three-dimensional histogram of met mast data.

https://doi.org/10.5194/amt-2019-200
Preprint. Discussion started: 11 September 2019

Atmospheric Measurement Techniques Discussions

Table 1. Most common atmospheric conditions identified by the three-dimensional histogram in Fig. 5

u [m/s]	θ [°]	TI [%]	Observations	Frequency [%]
[6.0, 9.0]	[280.0, 290.0]	[2.0, 4.0]	13376	0.67
[6.0, 9.0]	[290.0, 300.0]	[2.0, 4.0]	13343	0.65
[6.0, 9.0]	[270.0, 280.0]	[2.0, 4.0]	12255	0.61
[9.0, 12.0]	[290.0, 300.0]	[2.0, 4.0]	6423	0.59
[9.0, 12.0]	[290.0, 300.0]	[4.0, 6.0]	6096	0.56

Considering all three data channels simultaneously from an aggregate statistics perspective is accomplished with a three-dimensional volume-rendering of a histogram, shown in Fig. 5. Rendering of the three-dimensional histogram was accomplished with software produced by VAPOR (www.vapor.ucar.edu), a product of the Computational Information Systems Laboratory at the National Center for Atmospheric Research (Clyne et al., 2007; Clyne and Rast, 2005). The full histogram considers the interdependence of all three data channels together in a statistical sense and is often the means by which conditions of interest are identified. For model validation exercises, it is desirable to compare with statistically converged field observations. Thus, atmospheric conditions with good statistical representation are often selected as simulation or study cases. In the current data, the cases with the greatest representation correspond to wind speed, direction, and turbulence intensity in the ranges noted in Table 1. The number of observations reported for each atmospheric condition corresponds to the number of 1-min data points falling within the stated limits for u, θ, and TI. Color scale information provided in Fig. 5 reflects the interpolation undertaken in the generation of the three-dimensional histogram. The bins of atmospheric conditions reported in Table 1 are effectively larger than those in the three-dimensional histogram, hence, the reported observations are greater, as shown in Fig. 5. The frequencies reported in Table 1 represent observations in narrowly defined bins. Direct comparison to results reported above should take into account that observations are not integrated over other variables as in the one- or two-dimensional histograms.

4 Total variation of dynamical data

Aggregate statistical representation as in the three-dimensional histogram shown in Fig. 5 accounts for interdependence of the three variables considered in the current example, but cannot account for the dynamic nature of the atmosphere. A histogram, as a consequence of its composition, only denotes how frequently a given condition is observed without regard to what condition may precede or follow. For example, a given condition may be observed in any of the conditions noted in Table 1 only in a transient sense. The actual weather conditions could well be undergoing a dramatic change, but within any 1-min observation, the variables of interest fall within the stated bounds of a single bin within the full condition space.

An alternate path toward identifying conditions of interest for model validation or benchmarking studies comes through seeking continuous periods from the time series of observations that has properties of interest for a given study. An obvious choice would be a continuous period in which the atmospheric conditions remain steady or quasi-steady. A continuous time

https://doi.org/10.5194/amt-2019-200
Preprint. Discussion started: 11 September 2019

Atmospheric Measurement Techniques Discussions

series with quiescent conditions provides adequate convergence of measures of central tendency and of variability without sacrificing the information inherently contained in dynamically related observations. Additionally, retaining a time series allows users to leverage the interdependence of the channels within a data set by way of correlation or covariance metrics.

Quantifying the variability of a set of data must include the correlation between data channels, or risk discounting any information regarding the relationship between variables. Stated otherwise, any metric that combines the variability of each channel independently without accounting for covariance between the channels is incomplete and will not be sufficient to fully quantify or characterize the state of a given system. Therefore, a method that accounts for not only the variation within each channel, but also the interchannel variation is necessary to quantify the distribution of data across multiple channels into a single metric.

Below, each data block, $\mathbf{D}$, is a selected time period and corresponds to an array of size of $[m, n]$. In this case, m is the length of the time period — either 60 or 120 min —and n is three, corresponding to the number of variables u, θ, and TI.

$$\mathbf{D} = [u(t),\ \theta(t),\ TI(t)] \tag{1}$$

In addition to the definition of $\mathbf{D}$, a block, $\mathbf{f}$, containing objective functions of interest to apply to each of the variables in $\mathbf{D}$ is defined as,

$$\mathbf{f} = [f_u(t),\ f_\theta(t),\ f_{TI}(t)] \tag{2}$$

The difference between objective functions and their respective data is considered to be a regularized data block, and is noted with a caret,

$$\hat{\mathbf{D}} = \mathbf{D} - \mathbf{f} \tag{3}$$

The purpose of defining an objective function or set of functions is to tune the data to show covariance specifically with respect to a desired form about which the data are regularized. Seeking quiescent conditions in which minimal variation occurs in all data channels without regularization amounts to the special case of setting the function block to $\mathbf{f} = 0$ (or, more generally, when the objective function is any constant value; $\mathbf{f} = c$). The objective function block is discussed in greater detail in the following sections.

The total variation, $\mathcal{V}$, of a system is a unitless metric to quantify spread of a set of interdependent variables that accounts for autocorrelation within each channel and for covariance between channels. A covariance matrix is calculated for a subset taken from the full data, representing a continuous period of a specified duration,

$$\mathbf{C} = \hat{\mathbf{D}}^T \hat{\mathbf{D}} \tag{4}$$

$$= \begin{bmatrix} \sigma_u^2 & \sigma_u\sigma_\theta & \sigma_u\sigma_{TI} \\ \sigma_\theta\sigma_u & \sigma_\theta^2 & \sigma_\theta\sigma_{TI} \\ \sigma_{TI}\sigma_u & \sigma_{TI}\sigma_\theta & \sigma_{TI}^2 \end{bmatrix} \tag{5}$$

https://doi.org/10.5194/amt-2019-200
Preprint. Discussion started: 11 September 2019

Atmospheric Measurement Techniques Discussions

In Eq. (5), $\mathbf{C}$ is a square matrix of size $n \times n$ representing the covariance between any pair of data channels. The principal components of the covariance matrix are derived through an eigenvalue decomposition,

$$\mathbf{C}\mathbf{v} = \lambda\mathbf{v} \tag{6}$$

The eigenvectors are denoted as $\mathbf{v}$ and the eigenvalues as λ. By definition, the eigenvectors are a set of orthonormal vectors that most efficiently span the space of the covariance matrix. Principal components are eigenvectors weighted by their respective eigenvalues, $\mathcal{P} = \lambda\mathbf{v}$. Total variation, $\mathcal{V}$, is the vector summation of all principal components,

$$\mathcal{V} = \sum \mathcal{P} = ||\lambda|| \tag{7}$$

Given that the principal components are orthogonal, the total variation can be equivalently expressed as the L_2-norm of the eigenvalues.

4.1 Quiescent conditions: $\mathbf{f} = c$

In order for the variability of each channel in the data set, D, and their respective covariances to be given equal weight, the data must be normalized to a single range. In the current demonstration, each data channel has been normalized by its respective span and mapped to an interval determined by the range of each channel in standard deviations according to the formulation,

$$\mathbf{D}_{\text{norm}} = \frac{\mathbf{D} - \overline{\mathbf{D}}}{\sigma_{\mathbf{D}}} \tag{8}$$

In Eq. (8), the arithmetic mean and standard deviation (denoted by the overline and σ, respectively) are calculated separately for each column of $\mathbf{D}$. Normalizing data before calculating the total variation ensures that each data stream is weighted equally in the characterization of a given condition or state.

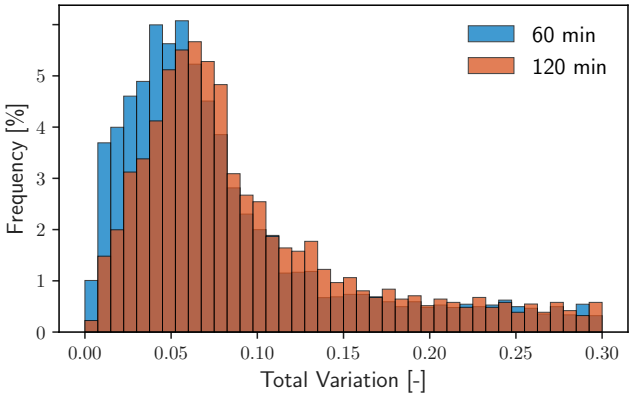


Figure 6. Distribution of $\mathcal{V}$ for data blocks of 60 or 120 min (blue and red, respectively)

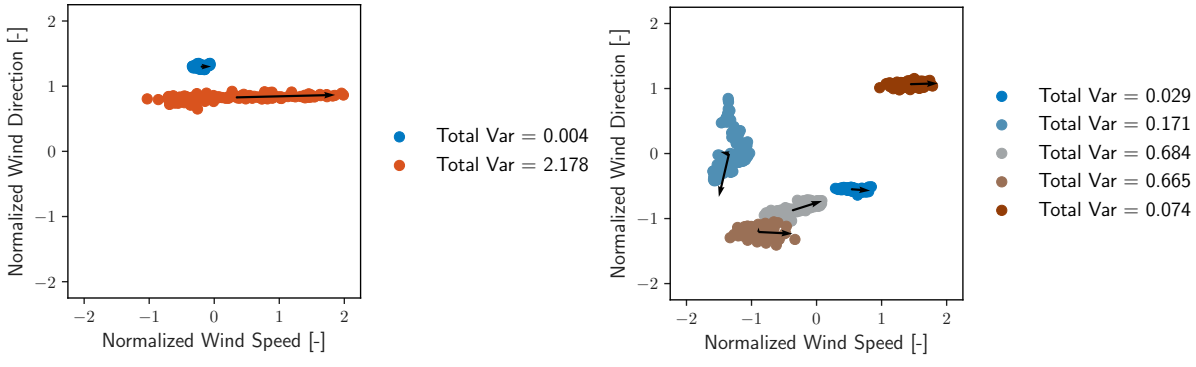
Fig. 6 shows the distribution of the total variation, $\mathcal{V}$ dividing the data record into periods of either 60 (blue) or 120 min (red). Immediately visible in the histograms of $\mathcal{V}$ is that there is a range of values exhibited most commonly by the blocks of

https://doi.org/10.5194/amt-2019-200
Preprint. Discussion started: 11 September 2019

Atmospheric Measurement Techniques Discussions

data. For data broken into 60-min periods, 35.9% of blocks have a total variation less than 0.05, whereas for data broken into 120-min periods, only 25.0% of blocks have a total variation in the same range. Although $\mathcal{V}$ is a unitless metric, its relative value does convey the degree of variation represented by all data within a respective time period. The values of $\mathcal{V}$ with the greatest frequency of occurrence is larger for periods of 120 min than for periods of 60 min. This is an expected trend because of the greater changes in atmospheric conditions that are possible within a larger window. There remains an inherent trade-off between the length of a data block and the degree of variation; longer blocks provide greater statistical convergence of $\mathbf{C}$, but risk including more dynamical variation, which contributes to higher values of $\mathcal{V}$.

Periods of time corresponding to the minimum values of $\mathcal{V}$ are those in which the total atmospheric conditions vary the least. In these periods, small values of standard deviation within each data channel as well as minimal covariance between the channels is expected. Minimal covariance between channels is equivalent to observing only stochastic, uncorrelated fluctuations in each channel. In contrast, periods corresponding to the maximum values of $\mathcal{V}$ are those in which the subset of data experiences the greatest variability, to which individual channel noise and correlated events between channels both contribute. To provide a sense of how other time periods are characterized in terms of $\mathcal{V}$, five randomly selected periods of 120 min are shown in Fig. 7(b). The principal components of each data block are shown with black vectors emanating from the center of each block and the total variation is listed in the legend. The figure represents each block of data as a scatter of only normalized wind speed and direction, although TI is also in the calculation of $\mathcal{V}$.



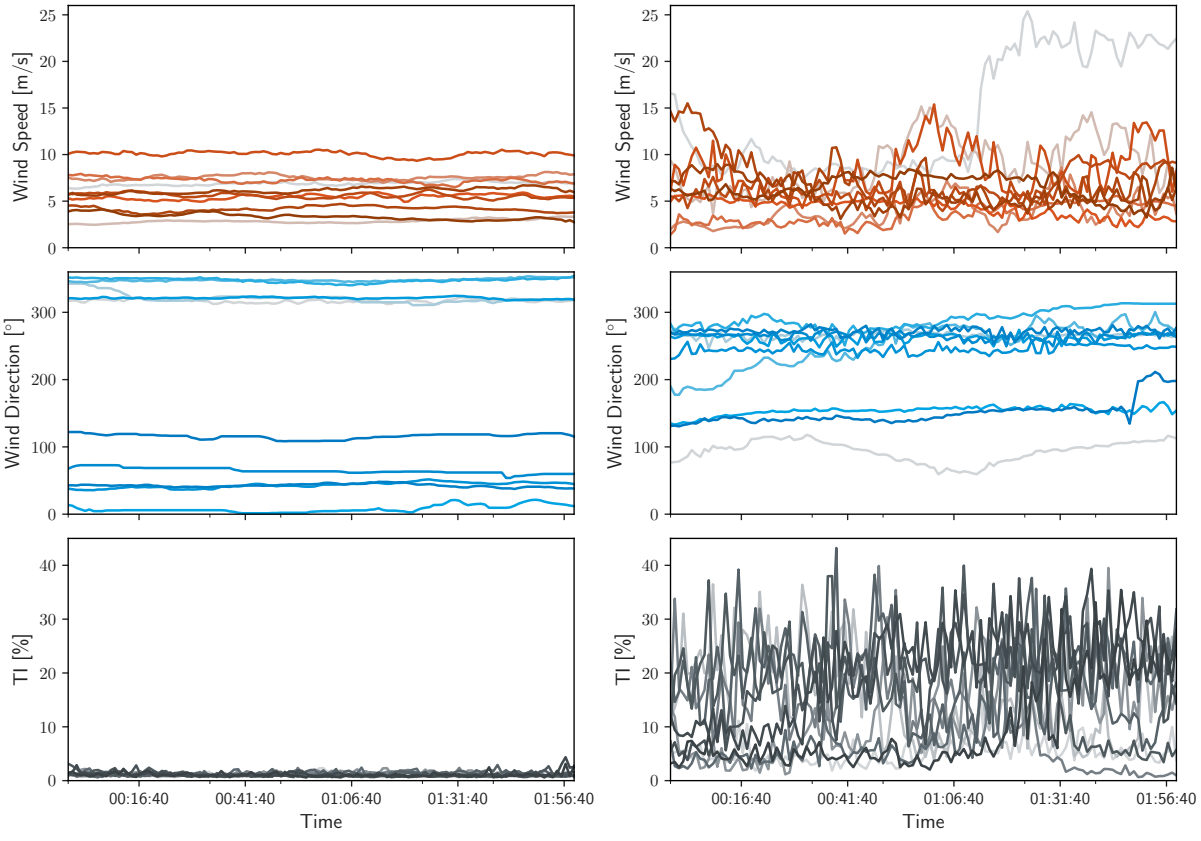
(a) Time periods exhibiting the lowest and highest values of total variation (blue and red, respectively)

(b) Scatter of observations in selected time periods

Figure 7. Scatter of data points of selected time periods within the full conditions space

Fig. 8 shows the wind speed, direction, and turbulence intensity corresponding to the 10 periods of minimum and maximum total variation. Each set of time series is shown in its original (non-normalized) engineering units to provide insight into the atmospheric conditions, although they were identified using normalized data. Fig. 8 shows that the periods with minimal values of $\mathcal{V}$ have time series that appear constant and experience only small stochastic variations within each channel and that periods with large values of $\mathcal{V}$ exhibit more spread. For each set of time series, the extreme values are shown in the boldest color (red, blue, and gray for the wind speed, direction, and turbulence intensity, respectively) and fade to lighter colors for more moderate

https://doi.org/10.5194/amt-2019-200
Preprint. Discussion started: 11 September 2019



(a) Time periods with minimum total variation

(b) Time periods with maximum total variation

Figure 8. Time series of the 10 blocks with minimum and maximum values of $\mathcal{V}$, (a) and (b), respectively

values of $\mathcal{V}$. Starting and ending times are not included, as Fig. 8 is intended only to demonstrate the sorting capability of the method.

4.2 Objective conditions: $f \neq 0$

Regularizing the data with respect to a set of nonzero objective functions centers the total variation calculation around specific conditions of interest. For example, in the case of wind plant analysis, it may be of interest to assess array performance during a wind speed ramp event or change of wind direction. Such events may be readily formulated according to accepted mathematical definitions and supplied to the total variation algorithm from Sect. 4. Defining specific objective functions will quantify the total system variability around those conditions, which can then be used to identify the time periods that match the event of interest most closely.

An additional step is considered to sort the full data set for a more general formulation. In such a case, events of interest are defined in a suitably general formulation, and a least-squares minimization is applied to seek the relevant parameter values.

https://doi.org/10.5194/amt-2019-200
Preprint. Discussion started: 11 September 2019

Atmospheric Measurement Techniques Discussions

In the current demonstration, function types of interest are wind speed ramps, wind speed waves, and wind direction changes, shown in Eq. (9), (10), and (11), respectively.

$$f_u(t) = c_0\, t + c_1 \tag{9}$$

$$f_u(t) = c_0 \sin\left(c_1\, t + c_2\right) + c_3 \tag{10}$$

$$f_\theta(t) = c_0 \arctan\left(c_1\, t + c_2\right) + c_3 \tag{11}$$

In each of the equations, objective function parameters, c_i, are sought through least-squares minimization. In the current case, the parameters, c_i, are chosen to minimize the following expressions,

$$\rho = \|\mathbf{D} - \mathbf{f}\|^2 \tag{12}$$

$$= \begin{cases} \min \sum \left(u(t) - f_u(t, c_i)\right)^2 \\ \min \sum \left(\theta(t) - f_\theta(t, c_i)\right)^2 \\ \min \sum \left(TI(t) - f_{TI}(t, c_i)\right)^2 \end{cases} \tag{13}$$

where ρ is the least-squares fit residual. Least-squares fit parameters and the respective fit residual from each time period are retained, enabling an additional layer of filtering for conditions of interest. After objective function coefficients are determined, the total variation method is continued, yielding a value of $\mathcal{V}$ for regularized data in each time period. Removal of the objective function amounts to "detrending" the data and determining the covariance of the remaining data.

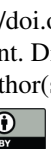
Fig. 9(a) compares distributions of $\mathcal{V}$ given the objective function definitions in Eq. (9), (10), and (11). The distributions indicate that the total variation can be reduced by regularizing data around generalized sinusoidal (red), linear (blue), and inverse tangent (black) functions as compared to the case where $\mathbf{f} = 0$ (gray). However, the reduction in $\mathcal{V}$ for the full data set is caused by the general definitions of the objective functions. Defining specific functions, even of the same forms, would likely increase the average value and spread of $\mathcal{V}$; it is not expected that a specific objective function would fit every time period well.

Noted earlier, the additional step of least-squares minimization provides a fit residual for each time period under consideration, shown in Fig. 9(b). Fit residuals indicate the goodness of fit of a given time period to the specified objective function forms. The distributions in Fig. 9(b) suggest that inverse tangent and sinusoidal functions fit the data with less residual error, ρ, than a linear objective function. This is likely caused by the additional objective function parameters (degrees of freedom) available for tuning the minimization.

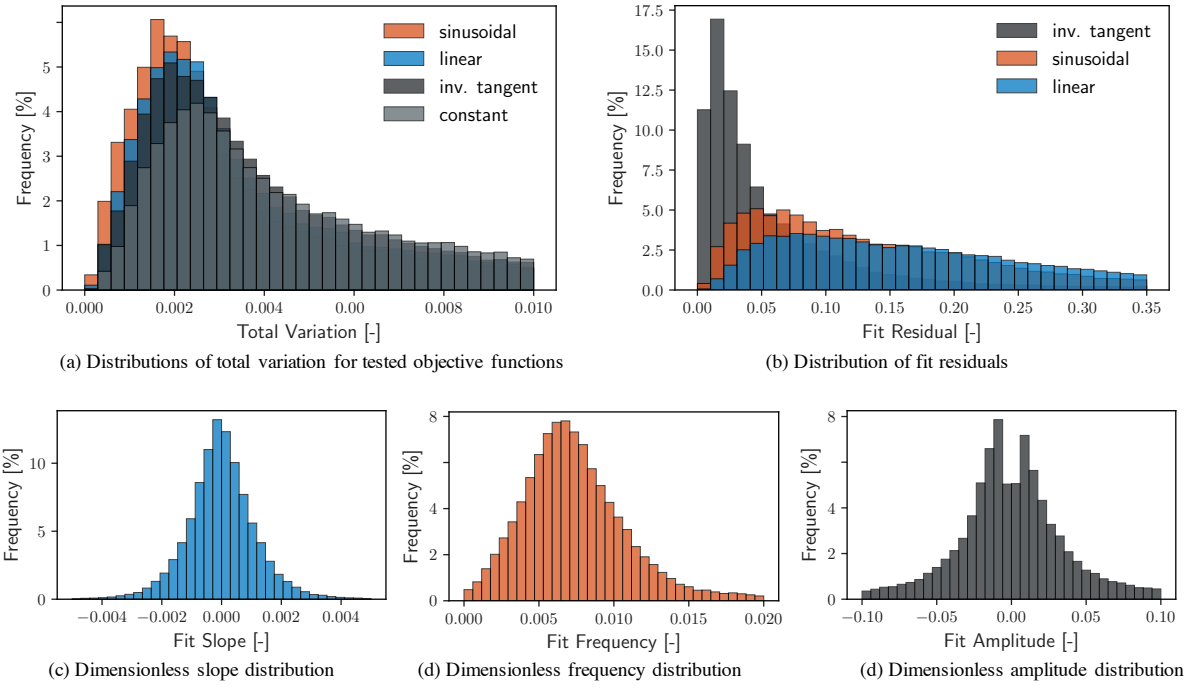
Adding an auxiliary step to the search process of least-squares minimization to a given objective function quantifies the goodness of fit of each data block and can return the parameter values necessary for the desired fit. For example, a least-squares fit to a linear relationship for any data channel will provide values of slope and offset as well as a residual value indicating the quality of the fit. In this way, the data provide alternative values for which sorting may be applied in addition to the total variation. As a demonstration, Figs. 10(a) and 10(b) show distributions of the best-fit slope to wind speed (the intensity of a speed ramp in (m/s)/min) and the fit residual (the goodness of fit to a linear objective function).

Figs 10(a) and 10(b) show a selection of periods with minimal total variation around linear and sinusoidal objective functions of wind speed, corresponding to wind speed ramps and waves, respectively. Selection of the wind speed ramps in Fig. 10(a)

https://doi.org/10.5194/amt-2019-200
Preprint. Discussion started: 11 September 2019



Atmospheric Measurement Techniques Discussions



(a) Distributions of total variation for tested objective functions

(b) Distribution of fit residuals

(c) Dimensionless slope distribution

(d) Dimensionless frequency distribution

(d) Dimensionless amplitude distribution

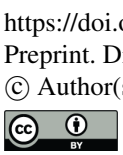
Figure 9. Distributions of selected quantities for selected objective functions

are conditioned to have the minimal total variation, minimal fit residual, and maximum absolute values of slope. These are the time periods in which the wind speed ramps are simultaneously the most well-behaved (i.e. minimal fit residual) and most intense (i.e. greatest absolute value of slope). Similarly, the wind speed waves shown in Fig. 10(b) were selected by seeking the minimal total variation and then selecting time periods in which the fit frequency fell between desired limits. In Fig. 10(b), the top subfigure shows 120-minute time periods in which the fit frequency is in the range of $[0.015, 0.02]$ rad/s (in red), and the bottom subfigure shows time periods in which the fit frequency is in the range of $[0.0075, 0.008]$ rad/s (in blue). Frequency limits were selected arbitrarily, and are meant only as a demonstration of the method's independence of fit frequency. Fig. 10(c) applies an inverse tangent objective function to the wind direction channel while seeking constant conditions in wind speed and turbulence intensity, identifying the periods of wind direction change with minimal total variation. Direction changes were considered in an absolute sense, and Fig. 10(c) shows time periods with minimal $\mathcal{V}$ in which the absolute direction change $|\Delta\theta|$ falls in the range $(20°, 40°]$. Again, the particular magnitude of direction change selected here is arbitrary, and was selected only to demonstrate the fit to an inverse tangent objective function.

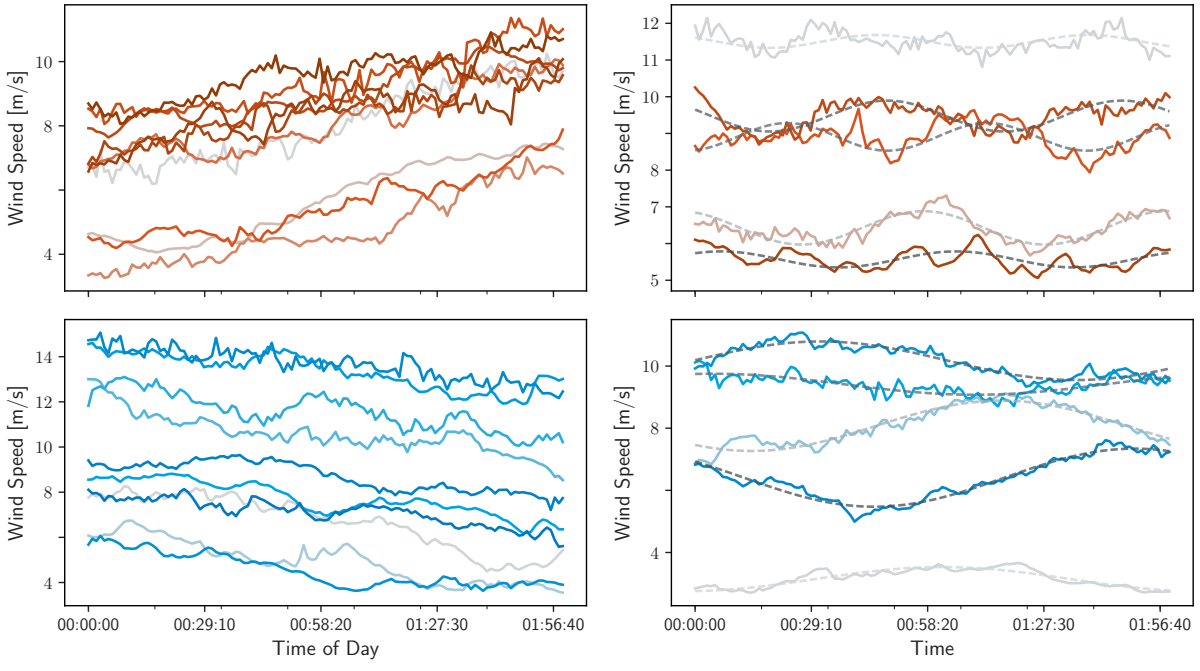
5 Sensitivity to outliers

A word of caution on using the total variation to identify periods of interest: Because principal component analysis is sensitive to outliers contained in the data, the method may falsely classify a time period as having a large value of total variation due to a

https://doi.org/10.5194/amt-2019-200
Preprint. Discussion started: 11 September 2019

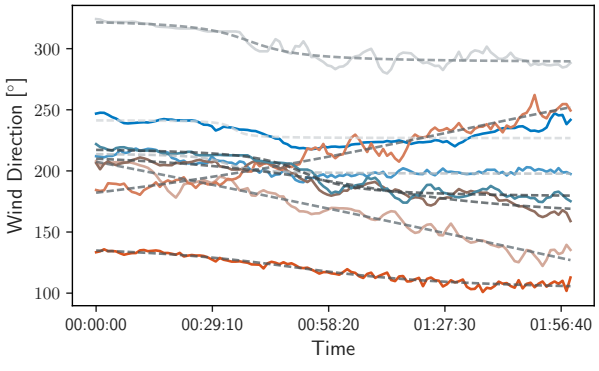


Atmospheric Measurement Techniques Discussions



(a) Most linear time series from met mast

(b) Wind speed wave time series from met mast



(c) Inverse tangent time series from met mast

Figure 10. Examples of time series identified by calculating covariance matrix around linear, sinusoidal, and inverse tangent objective functions

few spurious data points. Consideration of outliers in multivariate space requires a similar treatment as for the consideration of total variation. Seeking outlying points in each data channel individually discounts the possibility that the other data channels may be within acceptable statistical limits for the same point. Determining outliers from individual data channels further discounts any correlation that may exist between the channels. An effective means of considering outliers in multivariate data is the Mahalonbis distance, χ, which quantifies the Euclidean distance of a point from the center of a data set in terms of

https://doi.org/10.5194/amt-2019-200
Preprint. Discussion started: 11 September 2019

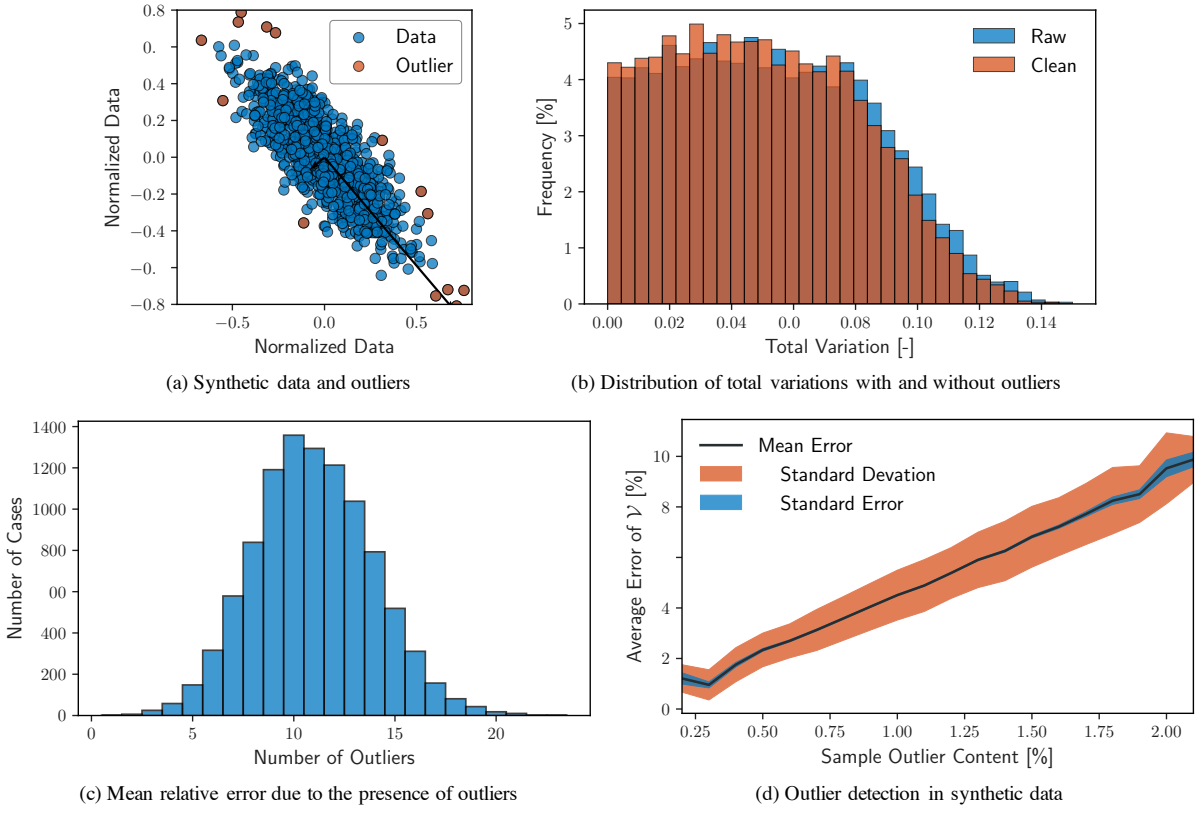
Atmospheric Measurement Techniques Discussions

standard deviations (De Maesschalck et al., 2000; Hadi, 1992; Rousseeuw and Van Zomeren, 1990; Xiang et al., 2008),

$$\chi = \sqrt{(x-\mu)^T C^{-1}(x-\mu)} \tag{14}$$

The Mahalonbis distance is sought through the covariance matrix of the data, and thus accounts for interdependence of the data channels, as emphasized earlier. Setting a threshold value for the Mahalonbis distance effectively draws an n-dimensional ellipsoidal boundary around the data set in nondimensional space, outside of which data are to be considered invalid.

To quantify the sensitivity of $\mathcal{V}$ to the presence of outliers, 10,000 synthetic data sets are generated, and outliers are detected and eliminated. Total variation is compared for each data set before and after outlier detection/elimination. Synthetic data sets (n=2 dimensions, 1,000 points each) are normally distributed about a zero mean value with a standard deviation that is randomly assigned in the range of [0, 10]. Each data set is normalized, given a random shape parameter to stretch the data, and rotated to simulate covariance between data channels. The covariance matrix is calculated using Eq. (5) and $\mathcal{V}$ calculated as in Eq. (7). Any point with $\chi > 3$ is flagged as an outlier and eliminated. The total variation is then calculated for the cleaned data.



(a) Synthetic data and outliers

(b) Distribution of total variations with and without outliers

(c) Mean relative error due to the presence of outliers

(d) Outlier detection in synthetic data

Figure 11. Outlier detection and the sensitivity of $\mathcal{V}$ to outliers

Fig. 11(a) shows a single example set of synthetic data. Accepted data are shown in blue, outliers in red, and the principal components of the data are shown as the black vectors. Fig. 11(b) shows distributions of $\mathcal{V}$ before and after exclusion of

https://doi.org/10.5194/amt-2019-200
Preprint. Discussion started: 11 September 2019


Atmospheric Measurement Techniques Discussions

outlying data identified with a threshold of χ in blue and red, respectively. As expected, the total variation of data sets without outliers is smaller than data sets before cleaning. Because of the large number of synthetic data sets considered, statistics regarding sensitivity to outliers are also within reach.

Fig. 11(c) shows the distribution of the number of detected outliers within each synthetic data set. Fig. 11(d) shows the mean relative error according to the number of detected outliers according to,

$$\varepsilon = \frac{\mathcal{V}_{\text{raw}} - \mathcal{V}_{\text{clean}}}{\mathcal{V}_{\text{raw}}} \tag{15}$$

where the subscripts denote the presence and absence of outliers as raw and clean, respectively. Uncertainty of the error is shown as the shaded bands around the mean relative error. The red band indicates the standard deviation of the relative error (σ_ε) and the blue band denotes the standard error ($\sigma_\varepsilon / N_{\text{outliers}}$). The roughly linear relationship shown in Fig. 11(d) indicates that one could expect an increase in error of approximately 4% for each additional percent outlier content of a given data set.

It should be noted that the present error analysis is not expected to yield identical results for atmospheric data. Observations of wind speed, direction, and turbulence intensity can vary considerably during any given period as part of the normal development of weather patterns. Mentioned briefly in the introduction, quality control of met mast and SCADA data is an active research topic and is beyond the scope of the current method development. However, it should be clear from the sensitivity analysis undertaken here that a careful quality control process should be applied before calculation of the total variation.

6 Conclusions

The definition of high-value conditions for wind plant analysis is ultimately up to the user, but may not conform to the most frequently observed state. For example, it may be of greater concern to wind plant developers, owners, or operators to be able to validate models where wake losses are greatest or during ramps of wind speed. These conditions may be more relevant to control or curtailment actions of wind plants, and may have a greater impact on the return on investment of wind energy assets.

Identification of continuous time periods that conform to conditions of interest is not intuitive through aggregate statistics, such as measures of central tendency or even joint probability distributions. The method to quantify the total variation of a multivariate data set described earlier provides a computationally economical means of parsing large and complex data sets, and includes a mathematically robust approach to sorting with respect to a desired condition or objective function. In addition, the method is independent of the length of the data record and of the number of channels included in the data sample. Normalizing the data makes combining disparate types of data into a single metric possible and meaningful.

The total variation method for seeking conditions of interest has applications far beyond the demonstration undertaken in the current work. Once properly classified, any number of detection and forecasting models may be trained and thoroughly validated. Collecting time periods containing similar dynamical events opens a path forward for more advanced analyses, such as modal decomposition methods and reduced order modeling. Extreme atmospheric events, as from the International Electrotechnical Commission (IEC) Standard for Wind Turbine Design (IEC, 2005), have well-defined characteristic functions and would thus fit well with the method explored in this article. After detection, wind turbine structural dynamics can be coupled to dynamical atmospheric events to produce robust and accurate control and cost models.

https://doi.org/10.5194/amt-2019-200
Preprint. Discussion started: 11 September 2019

Atmospheric Measurement Techniques Discussions

The total variation method explored here details identification and characterization of time series data from met masts only. Validation of high-fidelity wind plant models frequently resolves on some form of operational data, most often power production or some integrated statistic of wind plant performance. SCADA signals and power production or fault events could readily be identified with the total variation method. A further extension of the method would be to add functionality that accounts for spatial variation of operational data within a wind plant. A spatial aspect to the total variation method would augment the process to be able to detect and characterize the movement of weather fronts through a wind plant or cases in which wake losses are particularly significant and heterogeneous.

Acknowledgements. This work was authored by the National Renewable Energy Laboratory, operated by Alliance for Sustainable Energy, LLC, for the U.S. Department of Energy (DOE) under Contract No. DE-AC36-08GO28308. Funding provided by the U.S. Department of Energy Office of Energy Efficiency and Renewable Energy Wind Energy Technologies Office. The views expressed in the article do not necessarily represent the views of the DOE or the U.S. Government. The U.S. Government retains and the publisher, by accepting the article for publication, acknowledges that the U.S. Government retains a nonexclusive, paid-up, irrevocable, worldwide license to publish or reproduce the published form of this work, or allow others to do so, for U.S. Government purposes. Data was furnished to the authors under an agreement between the National Renewable Energy Laboratory, Siemens Gamesa Renewable Energy A/S, and Vattenfall. Data and results used herein do not reflect findings by Siemens Gamesa Renewable Energy A/S and Vattenfall. Additional thanks to the National Renewable Energy Laboratory Wake Squad for the musings and dialogue that ultimately led to this work getting started (and finished). Special thanks to Tony Martinez for countless discussions on everything from numerical methods to physical interpretations, editing, and computational assistance with volume rendering. Bridging the gap between my own turbulence experience and Mike Optis' atmospheric perspective essentially framed the discussion and motivation of the work.

https://doi.org/10.5194/amt-2019-200
Preprint. Discussion started: 11 September 2019

Atmospheric Measurement Techniques Discussions

References

Ali, N., Hamilton, N., Calaf, M., and Cal, R. B.: Turbulence kinetic energy budget and conditional sampling of momentum, scalar, and intermittency fluxes in thermally stratified wind farms, Journal of Turbulence, pp. 1–32, 2019.

Barthelmie, R., Crippa, P., Wang, H., Smith, C., Krishnamurthy, R., Choukulkar, A., Calhoun, R., Valyou, D., Marzocca, P., Matthiesen, D., et al.: 3D wind and turbulence characteristics of the atmospheric boundary layer, Bulletin of the American Meteorological Society, 95, 743–756, 2014.

Barthelmie, R., Churchfield, M. J., Moriarty, P. J., Lundquist, J. K., Oxley, G., Hahn, S., and Pryor, S.: The role of atmospheric stability/turbulence on wakes at the Egmond aan Zee offshore wind farm, in: Journal of Physics: Conference Series, vol. 625, p. 012002, IOP Publishing, 2015.

Belušić, D. and Mahrt, L.: Is geometry more universal than physics in atmospheric boundary layer flow?, Journal of Geophysical Research: Atmospheres, 117, 2012.

Bossavy, A., Girard, R., and Kariniotakis, G.: Forecasting ramps of wind power production with numerical weather prediction ensembles, Wind Energy, 16, 51–63, 2013.

Clyne, J. and Rast, M.: A prototype discovery environment for analyzing and visualizing terascale turbulent fluid flow simulations, in: Electronic Imaging 2005, pp. 284–294, International Society for Optics and Photonics, 2005.

Clyne, J., Mininni, P., Norton, A., and Rast, M.: Interactive desktop analysis of high resolution simulations: application to turbulent plume dynamics and current sheet formation, New Journal of Physics, 9, 301, 2007.

Cutler, N., Kay, M., Jacka, K., and Nielsen, T. S.: Detecting, categorizing and forecasting large ramps in wind farm power output using meteorological observations and WPPT, Wind Energy: An International Journal for Progress and Applications in Wind Power Conversion Technology, 10, 453–470, 2007.

De Maesschalck, R., Jouan-Rimbaud, D., and Massart, D. L.: The mahalanobis distance, Chemometrics and intelligent laboratory systems, 50, 1–18, 2000.

Eaton, M. L.: Multivariate statistics: a vector space approach., JOHN WILEY & SONS, INC., 605 THIRD AVE., NEW YORK, NY 10158, USA, 1983, 512, 1983.

Ferreira, C., Gama, J., Matias, L., Botterud, A., and Wang, J.: A survey on wind power ramp forecasting., Tech. rep., Argonne National Lab.(ANL), Argonne, IL (United States), 2011.

Ferreira, C., Gama, J., Miranda, V., and Botterud, A.: Probabilistic ramp detection and forecasting for wind power prediction, in: Reliability and risk evaluation of wind integrated power systems, pp. 29–44, Springer, 2013.

Fulcher, B. D.: Feature-based time-series analysis, in: Feature Engineering for Machine Learning and Data Analytics, pp. 87–116, CRC Press, 2018.

Gamage, N. and Hagelberg, C.: Detection and analysis of microfronts and associated coherent events using localized transforms, Journal of the atmospheric sciences, 50, 750–756, 1993.

Hadi, A. S.: Identifying multiple outliers in multivariate data, Journal of the Royal Statistical Society: Series B (Methodological), 54, 761–771, 1992.

Hannesdóttir, A. and Kelly, M.: Detection and characterization of extreme wind speed ramps, Wind Energy Science Discussions, 2019, 1–18, 2019.

https://doi.org/10.5194/amt-2019-200
Preprint. Discussion started: 11 September 2019

Atmospheric Measurement Techniques Discussions

Hansen, K. S., Barthelmie, R. J., Jensen, L. E., and Sommer, A.: The impact of turbulence intensity and atmospheric stability on power deficits due to wind turbine wakes at Horns Rev wind farm, Wind Energy, 15, 183–196, 2012.

IEC, I.: 61400-1: Wind turbines part 1: Design requirements, International Electrotechnical Commission, p. 177, 2005.

Kang, Y., Belušić, D., and Smith-Miles, K.: Detecting and classifying events in noisy time series, Journal of the Atmospheric Sciences, 71, 1090–1104, 2014.

Kang, Y., Hyndman, R. J., and Smith-Miles, K.: Visualising forecasting algorithm performance using time series instance spaces, International Journal of Forecasting, 33, 345–358, 2017.

Kumar, P. and Foufoula-Georgiou, E.: Wavelet analysis for geophysical applications, Reviews of geophysics, 35, 385–412, 1997.

Kumar, V., Kleissl, J., Meneveau, C., and Parlange, M. B.: Large-eddy simulation of a diurnal cycle of the atmospheric boundary layer: Atmospheric stability and scaling issues, Water resources research, 42, 2006.

Larsen, T. J., Madsen, H. A., Larsen, G. C., and Hansen, K. S.: Validation of the dynamic wake meander model for loads and power production in the Egmond aan Zee wind farm, Wind Energy, 16, 605–624, 2013.

Lilly, J. M.: Element analysis: a wavelet-based method for analysing time-localized events in noisy time series, Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences, 473, 20160 776, 2017.

Preston, D., Protopapas, P., and Brodley, C.: Discovering arbitrary event types in time series, Statistical Analysis and Data Mining: The ASA Data Science Journal, 2, 396–411, 2009.

Rousseeuw, P. J. and Van Zomeren, B. C.: Unmasking multivariate outliers and leverage points, Journal of the American Statistical association, 85, 633–639, 1990.

Sevlian, R. and Rajagopal, R.: Wind power ramps: Detection and statistics, in: 2012 IEEE Power and Energy Society General Meeting, pp. 1–8, IEEE, 2012.

Shahabi, C. and Yan, D.: Real-time Pattern Isolation and Recognition Over Immersive Sensor Data Streams., in: MMM, pp. 93–113, 2003.

Sørensen, J. N. and Shen, W. Z.: Numerical modeling of wind turbine wakes, Journal of fluids engineering, 124, 393–399, 2002.

Sun, J., Nappo, C. J., Mahrt, L., Belušić, D., Grisogono, B., Stauffer, D. R., Pulido, M., Staquet, C., Jiang, Q., Pouquet, A., et al.: Review of wave-turbulence interactions in the stable atmospheric boundary layer, Reviews of geophysics, 53, 956–993, 2015.

Taswell, C.: Handbook of wavelet transform algorithms, 2001.

Wasserman, L.: All of statistics: a concise course in statistical inference, Springer Science & Business Media, 2013.

Xiang, S., Nie, F., and Zhang, C.: Learning a Mahalanobis distance metric for data clustering and classification, Pattern recognition, 41, 3600–3612, 2008.

Zhang, J., Florita, A., Hodge, B.-M., and Freedman, J.: Ramp forecasting performance from improved short-term wind power forecasting, in: ASME 2014 international design engineering technical conferences and computers and information in engineering conference, pp. V02AT03A022–V02AT03A022, American Society of Mechanical Engineers, 2014.